(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 008 572 A1

(12) **EUROPEAN PATENT APPLICATION**
published in accordance with Art. 158(3) EPC

(43) Date of publication:
14.06.2000 Bulletin 2000/24

(21) Application number: **98929806.2**

(22) Date of filing: **02.07.1998**

(51) Int. Cl.[7]: **C07B 61/00**, G06F 17/50, G06F 17/30, C07K 1/00

(86) International application number:
PCT/JP98/02986

(87) International publication number:
WO 99/01409 (14.01.1999 Gazette 1999/02)

(84) Designated Contracting States:
CH DE FR GB IT LI

(30) Priority: **03.07.1997 JP 17793397**

(71) Applicant: **ITAI, Akiko**
Tokyo 113-0033 (JP)

(72) Inventors:
- **ITAI, Akiko**
  Bunkyo-ku, Tokyo 113-0033 (JP)
- **IMAMURA, Masazumi,**
  Room 301, Garden-Court
  Chiba-shi, Chiba 267-0066 (JP)

(74) Representative:
Nicholls, Michael John et al
J.A. KEMP & CO.
14, South Square
Gray's Inn
London WC1R 5LX (GB)

(54) METHOD FOR INFERRING PROTEIN FUNCTIONS WITH THE USE OF LIGAND DATA BASE

(57) A method of predicting biological functions of query proteins whose steric structures are known or predictable, using a three-dimensional structure database storing bio-active compounds which bind to target proteins with known biological functions, which comprises the steps of:

(1) extracting bio-active compounds capable of binding to said query protein as ligand candidates from said database based on the capability of complex formation between the query protein and bio-active compounds; and
(2) predicting that biological functions of the query protein are identical or similar to the biological functions of the target protein to which said ligand candidates bind.

EP 1 008 572 A1

Description

Technical Field

[0001] The present invention relates to prediction methods of the protein functions and databases used for said methods.

Background Art

[0002] Proteins are biopolymers comprising 20 kinds of amino acids as building blocks and have structures in which about 50 to 1,000 amino acids are connected in a chain by peptide bonds (-CONH-). The existence of various kinds of proteins has been revealed such as enzymes which catalyze substance conversion in organism, receptors related to their inter or intracellular signal transduction, receptors related to the control of gene expression, cytokines which are secreted at the time of inflammation, proteins related to the transport of substances and others. In the organisms of higher animals such as human, there are 50 to 100 thousands of kinds of proteins, and each plays specific functions and roles.

[0003] Enzymes provide fields for chemical reactions in which specific products are obtained by the actions on specific substrates, and proceed stereospecific or regiospecific reactions with moderate conditions. Receptors transduce signals through the structural change upon the binding of hormones and signal transmitters. The features common to these enzymes and receptors are the appearance of their biological functions by forming stable complexes with specific molecules (ligands). Protein molecules, which are long like strings, are folded to take certain steric structures and form structural sites (ligand binding sites) which bind specifically with artificial molecules such as drugs and specific biomolecules. This ligand binding site is essential for the appearances of the functions of enzymes and receptors.

[0004] The steric structures of proteins can be determined by X-ray crystallographic analysis and NMR analysis. Due to the remarkable progress and spread of these analytical techniques, determination of steric structures of proteins has become easy, and the number of proteins analyzed is increasing acceleratingly. Protein Data Bank, which is a database of protein structures, stores three-dimensional coordinates of more than 7,000 proteins at present, and the data are available throughout the world. Accordingly, once functions of a protein are known, it has become possible to understand the relations between the structure and the function of the protein on atomic levels by analyzing the crystal structure of the complexes with appropriate ligands. Moreover, by using the steric structures of proteins which have been analyzed crystallographically as templates, and by substituting the side chains of amino adds, it has become possible to predict the steric structure of a protein having highly homologous amino acid sequences (homology modeling).

[0005] Protein studies have so far been conducted by the means in which after the separation and purification of proteins employing its biological function as a guide, its amino add sequence is determined to analyze the structure and function. However, recently, as analyses of genes have become easy, there are cases in which the existence of a protein is suggested from genetic information. For example, the existence of considerable number of proteins has been revealed by a large-scale project aiming at the human genome analysis, and these results are expected to be utilized for the elucidation of the cause of diseases and drug design.

[0006] However, for those proteins successively found from genome analysis studies, their amino acid sequences are merely elucidated, while in most cases their biological functions cannot be predicted at all. For this reason, an enormous amount of study is necessary to predict or confirm functions for each protein, which becomes an obstacle for the effective use of genome information. Moreover, although the steric structure of proteins whose amino acid sequences have been elucidated can be determined more easily than before due to the progress of crystallographic analysis and NMR analysis, there are many cases in which the functions are hardly known even though the steric structures of proteins have been elucidated.

[0007] At present, methods of predicting the functions of novel proteins easily have not been established. For example, a prediction method is adopted in which a novel protein is predicted to have functions similar to a known protein, if a protein with high homology is found by comparing the amino acid sequence of the novel protein with groups of amino acid sequences of proteins with known functions. Furthermore, for the multiple proteins with the same functions, information concerning the correlation between the structure and function can be obtained by making alignment so that homologous parts become as large as possible. However, even for proteins with the same function, the homology is not so high in general when the biological species are different. Thus, the above-mentioned methods which depend on alignment are not helpful at all for many proteins whose functions are known to be the same or not.

Disclosure of Invention

[0008] An object of the present invention is to provide methods of predicting functions of proteins. More specifically, the object of the present invention is to provide methods to predict easily the functions and roles in organism, for pro-

teins whose steric structures are known or predictable. Moreover, another subject of the present invention is to provide a database which is helpful for exploring shapes and properties of ligand binding site of proteins from the side of bio-active compounds (ligands).

[0009] As a result of zealous endeavor to solve above-mentioned subjects, the inventors of the present invention found that the functions of proteins without known functions can be predicted with good accuracy by preparing a three-dimensional structure database which stores bio-active compounds capable of binding as ligands with target proteins with known biological functions, judging capability of complex formation between the proteins without known functions and each bio-active compound in the database, and selecting bio-active compounds with high capability of complex formation as ligand candidates. The present invention was achieved based on these findings.

[0010] The present invention thus provides methods of predicting biological functions of query proteins whose steric structures are known or predictable, using a three-dimensional structure database which stores one or more bio-active compounds which bind to target proteins with known biological functions, which comprises the steps of:

(1) extracting bio-active compounds capable of binding to said query protein from said database as ligand candidates, based on the capability of complex formation between the query protein and the bio-active compounds; and
(2) predicting that biological functions of the query protein are identical or similar to the biological functions of the target proteins to which said ligand candidates bind. According to a preferred embodiment of the present invention, the above-mentioned method comprises the steps of:
(3) extracting one or more ligand binding sites for the query protein;
(4) exploring the most stable complex formed with the ligand binding sites of the query protein for each bio-active compound included in the database;
(5) extracting bio-active compounds which satisfy hit conditions preset, based on the stabilities and structural features of the most stable complexes;
(6) extracting further, as required, bio-active compounds from the bio-active compounds extracted in step (5) which satisfy hit conditions different from those in the above-mentioned step (5); and
(7) predicting that biological functions of the query protein are identical or similar to the biological functions of the target protein to which said ligand candidates bind, while treating the bio-active compounds extracted in above-mentioned steps (5) or (6) as ligand candidates.

[0011] In further preferable methods of the present invention, above-mentioned steps (4) through (6) are performed automatically using the program ADAM&EVE (PCT/JP95/02219: WO96/13785). According to other embodiments of the present invention, there are provided a method of predicting biological functions of query proteins using a three-dimensional database which stores one or more bio-active compounds which bind to target proteins with known biological functions; a method of predicting biological functions of query proteins by exploring the shapes and properties of the ligand binding sites of the query protein using one or more bio-active compounds which bind to target proteins with known biological functions which are stored in a three-dimensional database; and a method of predicting functions of query proteins by extracting ligand candidates for the query protein from a three-dimensional database which stores one or more bio-active compounds which bind to target proteins with known biological functions.

[0012] According to still other embodiment of the present invention, there is provided a three-dimensional database which stores one or more bio-active compounds which bind to target proteins with known biological functions and is used for each of the above-mentioned methods. According a preferred embodiment of the present invention, there is provided a database including information about the target protein for each bio-active compound, and in a further preferred embodiment, the above-mentioned database is prepared in a form which enables to perform the above-mentioned steps (4) through (6) automatically using the program ADAM&EVE (in the specification, the database is sometimes referred to as "ADAM-style database").

[0013] From other points of view, the present invention provides a method of predicting biological functions of query proteins whose steric structures are known or predictable, using a three-dimensional database which stores one or more intrinsic bio-active compounds with known bio-activities in organisms but without the knowledge of the target protein, which comprises the steps of:

(1) extracting bio-active compounds capable of binding to said query protein from said database as ligand candidates, based on the capability of complex formation between the query protein and bio-active compounds; and
(8) predicting that biological functions of the query protein concern the bio-activities of said ligand.

[0014] According to a preferred embodiment of this method, there is provided the above-mentioned method comprising steps of:

(3) extracting one or more ligand binding sites for the query protein;

(4) exploring the most stable complex formed with the ligand binding sites of the query protein for each bio-active compound included in the database;

(5) extracting bio-active compounds which satisfy hit conditions preset, based on the stabilities and structural features of the most stable complexes;

(6) further extracting, as required, bio-active compounds from the bio-active compounds extracted in step (5) which satisfy hit conditions different from those in the above-mentioned step (5); and

(9) predicting that biological functions of the query protein concern the bio-activity of the ligand, while treating the bio-active compounds capable of complex formation extracted in steps (5) or (6) as ligand candidates.

[0015] As a more preferred embodiment, there is provided the above-mentioned method in which steps (4) through (6) are performed automatically using the program ADAM&EVE.

[0016] Furthermore, there are provided by the present invention, a method of predicting biological functions of query proteins whose steric structures are known or predictable using a three-dimensional database which stores one or more intrinsic bio-active compounds with known bio-activities in organisms but without the knowledge of the target protein; a method of predicting biological functions of query proteins by exploring the shapes and properties of the ligand binding site of the query protein using one or more intrinsic bio-active compounds with known bio-activities in organisms but without the knowledge of the target protein stored in a three-dimensional database; and a method of predicting biological functions of query proteins by extracting ligand candidates for the query protein from a three-dimensional database which stores one or more intrinsic bio-active compounds with known bio-activities in organisms but without the knowledge of the target protein.

[0017] In addition, the present invention provides a three-dimensional database which stores one or more intrinsic bio-active compounds with known bio-activities in organisms but without the knowledge of the target protein, and which is used for the above-mentioned methods. According to a preferred embodiment of the present invention, there is provided a database including the information about bio-activity of each bio-active compound, and according to a more preferred embodiment, the above-mentioned database is prepared in a form which enables to perform the above-mentioned steps (4) through (6) automatically using the program ADAM&EVE.

Brief Explanation of Drawing

[0018]

Figure 1 shows the three-dimensional structure and ligand binding sites of bovine trypsin.

Figure 2 shows the binding mode of nafamostat extracted from the database as a ligand candidate.

Most Preferred Embodiment for Carrying Out the Invention

[Preparation of Database]

[0019] For carrying out the methods of the present invention, it is preferable to prepare in advance a three-dimensional database which stores bio-active compounds which bind as ligands to target proteins with known biological functions. The kinds of bio-active compounds are not particularly limited, for example, various bio-active substances which exist in organism, for example, transmitters (receptor substrate), enzyme substrates and enzyme products, vitamins, hormones, autacoids, co-enzymes, amino acids, bio-active peptides, nucleotides, monosaccharides in glycolytic pathway, or organic acids, as well as medicinal molecules, enzyme inhibitors, or toxins, which do not originally exist in organism, may be acceptable. Moreover, not only substances with low molecular weight but also compounds with high molecular weight such as proteins, nucleic acids like RNA or DNA, or polysaccharides may be acceptable.

[0020] In order to increase the accuracy of prediction of the methods of the present invention, it is desirable to store a lot of bio-active compounds in the above-mentioned database so that diverse biological functions are covered. It is also desirable to store as many bio-active compounds as possible with various molecular skeletons for each bio-activity, although it is sufficient to store at least one typical bio-active compound for one bio-activity in the database. Furthermore, one may prepare more than two kinds of appropriate databases, and select a desirable database and use it for the methods of the present invention.

[0021] For each bio-active compound stored in the database, it is desirable to store additional information such as information about the structure of compounds; information about the binding with target proteins; information about the biological functions of target proteins; information about the bio-activity of the compounds in case the target protein is unknown. Examples of these information include one or more kinds of information selected from the following group, comprising: name of the compound; number of constituting atoms and molecular weight; element name of each atom; two-dimensional coordinates and three-dimensional coordinates atom types for force-field calculation; atomic charge;

bonding relations; modeling method; conformation role; bio-activity; name of target protein; subunit or domain; function classification biological species binding constant or sub-type specificity; and steric structure information.

[0022] In the database of the present invention, it is preferable that all of these are added as information. However, information about bio-active substances is not limited to the above-mentioned items, and one or more of the items may be substituted by other information. Furthermore, other information may be added to the above-mentioned information as required. Concerning these information, it is not always necessary to store them in a single database, and it is acceptable as long as some relationships are retained, for example, by including tag information which points to record or data in each database. In the following, each information is explained more specifically. However, it should be understood that these are explained as examples and persons skilled in the art can appropriately select them.

[0023] As "compound name", any names such as common name, trade name, development code, IUPAC nomenclature may be used so long as it can identify the bio-active compound. "Number of constituting atoms" is the number of each constituting element included in the bio-active compound, which may be expressed, for example, like $C_{24}H_{20}O_2$. For "three-dimensional coordinates", those expressed by orthogonal axes (x, y, z) in angstrom unit are preferable. "Atom types for force-field calculation" mean symbols or numbers for further classifying elements based on orbital hybridization and the like, which are used for calculating force-field energy. "Atomic charge" is the formal charge assigned on each atom to calculate electrostatic interaction energy in the force-field energy and "bonding relation" is the information showing which atom forms a covalent bond with which number atom in the molecule, and how many the order of the bond is.

[0024] "Modeling method" is an information which indicates the origin of the three-dimensional structure of bio-active compounds, and includes information such as whether the three-dimensional structure is derived from a sole crystal structure or whether the three-dimensional structure is predicted from the two-dimensional structure using a program which convert to three dimensions. "Conformation" indicates information such as whether the conformation of the three-dimensional structure is one of the local-minimum structures obtained from the three-dimensional conversion, a sole crystal structure, or a structure determined by NMR, and whether or not the conformation is active conformation. "Role" is an information which shows that the bio-active compound acts as, for the target protein, which of enzyme substrate, enzyme reaction product, enzyme inhibitor, co-enzyme, effector, intrinsic ligand, agonist, antagonist, receptor substrate and the like. "Bio-activity" is an information about the change caused in organism upon administration of the bio-active compound.

[0025] For "target protein name", it is preferable to adopt those which generally include the functions of the protein, for example, dihydrofolate reductase, retinoid receptor and others. "Subunit or domain" is an information which indicates the subunit or domain to which bio-active compounds bind when the target protein consists of multiple subunits or domains. "Function classification" means a broad classification of function of the target protein in organism, which is exemplified by information including classification such as enzyme, trans-membrane receptor, nuclear receptor, cytokine, and transporter protein.

[0026] Information of "biological species" includes information about the biological species from which the target protein is derived. It is usually specified by taxonomy by species, genus, family, class and the like. More practically, one may use classification such that 1 for all biological species, 2 for higher animals, 3 for lower animals, 4 for prokaryotes, and 5 for plants. "Tissue" may at least include information about tissues where the target protein mainly exist and is functioning, which is exemplified by tissue names such as blood, liver and others for the case of human species. "Binding constant and sub-type specificity" may include information such as binding constant, IC_{50}, sub-type specificity of binding. Information about "steric structure" includes information whether three-dimensional structure of the target protein is known or not, and it is desirable to include information about whether the analytical method is crystallographic analysis or NMR analysis when three-dimensional structure is known and to include the code number in the Protein Data Bank if the structure is available therefrom.

[Extraction of Ligand Binding Site on Query Protein]

[0027] Concerning query proteins for which biological functions are to be predicted, there is no limitation about their kinds or sizes as long as their steric structures are known or predictable. For example, proteins consisting of multiple subunits or conjugated proteins like glycoprotein may be acceptable. If the three-dimensional structure analysis has been performed for the query protein by crystallographic or NMR method, data on the steric structure can be used directly. Alternatively, the steric structure may be predicted by homology modeling method and the like, using steric structures of homologous proteins as templates.

[0028] In order to judge capability of complex formation between the query protein and bio-active compounds in the database, one or more sites in the query protein molecule are extracted as candidates for the ligand binding sites. Generally, this step may be performed interactively by rotating the query protein molecule on computer graphics display and judging visually the sizes and depths of the sites like a pocket or a cavity on the molecular surface that have characteristic shapes and properties to be ligand binding sites. Alternatively, it is also possible to explore these sites automati-

cally. If more than two candidate sites are found on the molecular surface of the query protein, the following exploration steps may be performed regarding each as a ligand binding site.

[Exploration of Stable Complex Formed between Bio-active Compound and Query Protein]

[0029] Judgment of capability of complex formation is conducted between one or more ligand binding sites found in the query protein and each bio-active compound stored in the database. Capability of complex formation may be judged, for example, based on the stability (such as low energy value) and structural features of the complex after forming one or more complexes by binding one of the bio-active compounds stored in the database to the ligand binding site on the query protein. In order to explore multiple stable complexes effectively which are formed between the bio-active compounds and the ligand binding site of the query protein, a simulation called docking study may be utilized.

[0030] This method generally includes a process of displaying the ligand binding sites of a protein with a known structure on computer graphics display, and a process of exploring locations for stable binding by rotating and translating the molecule to be bound, with these processes usually conducted interactively. For the molecules with flexible conformation which have rotatable bonds, it is preferable to include a process of exploring stable locations while varying the conformation. After obtaining several locations which may lead to stable binding, it is possible to predict the most stable structure of the complex by performing energy calculation and optimization as required.

[0031] As a programs for the docking study, a program developed by Tomioka and others (GREEN) may be suitably employed, for example (Tomioka, N. and Itai, A., J. Comput.-Aided Mol. Design, 8, pp.347-366, 1994). However, since a freedom concerning the rotation and translation of molecules and a freedom of conformation are coupled together, there are cases in which the above-mentioned interactive method is not sufficient for predicting the most stable structure with comprehension of all possible combinations. As a method of exploring the most stable structure of complexes while solving such problems, a program by Mizutani and others (ADAM), which performs docking automatically, can be suitably employed(Mizutani, Y.M. et al., J. Mol. Biol., 243, pp.310-326, 1994; US Patent No. 5,642,292; PCT/JP93/0365).

[0032] When the program ADAM is employed, it is possible to explore several to several dozens of stable complex structures including the structure of the most stable complex effectively out of tremendous amount of complex structures resulting from the freedoms of binding mode and conformation, and it is possible to output automatically the complex structures obtained from the exploration, sorted in an order of their stabilities and other indices. The program ADAM, whose characteristics is high reliability and accuracy, includes a process of structure optimization with location and torsion angles varied continuously by means of repeated energy minimization, which is conducted after comprehension of approximate possibility of bonding mode and ligand conformation based on the geometrical condition of hydrogen bond formation.

[0033] In order to predict complex structures using the program ADAM, it is generally necessary to specify atom-type number and atomic charge of each atom of the bio-active compound, which is used for force-field energy calculation, classification number of hydrogen bonding functional groups for heteroatoms, initial value, final value, and increment value of torsion angle for rotatable bonds, which is used for generation of conformation, as well as the three-dimensional coordinates of the query protein and the bio-active compound. These parameters can be input interactively on computer graphics display when a bio-active compound included in the database is processed one by one using the program ADAM.

[Extraction of Bio-active Compounds to be Ligand Candidates]

[0034] By evaluating capability of complex formation between each bio-active compound and the query protein, it is possible to extract compounds that can bind to the query protein stably as ligand candidates, out of the bio-active compounds stored in the database. In the most preferable embodiment, the above-mentioned exploration process of complex structures and extraction process of ligand candidates may be conducted automatically in a consecutive process using the program ADAM&EVE (PCT/JP95/02219:WO96/13785).

[0035] When the program ADAM&EVE is employed, only a complex which is most stable (most stable complex) is explored automatically out of the complexes formed with the query protein, for each of the diverse and a lot of numbers of bio-active compounds stored in the database. After that, a judgement is given to that most stable complex whether it satisfies the criteria of selection (hit conditions) preset, and then one or more bio-active compounds satisfying the criteria are extracted as ligand candidates. As for hit conditions, parameters regarding the stabilities of the complexes (energy values) and regarding the structural features may be generally adopted. For example, value of intermolecular interaction energy, number of hydrogen bonds, molecular weight, number of atoms, number of rings, ionic bonding or hydrogen bonding with specific functional groups in the proteins may be specified arbitrarily.

[0036] When the exploration process of complex structures and extraction process of ligand candidates are conducted by the program ADAM&EVE, it is desirable to include in the database, coordinates of hydrogen atoms, atom-

type number and atomic charge for each atom of bio-active compound, which is used for the force-field energy calculation, classification number of hydrogen-bonding functional groups for heteroatoms, rotatable bond and the information on their rotation (initial value, final value, increment value of torsion angle) and the like, as well as the three-dimensional coordinates of the query protein and the bio-active compound, so that diverse and many numbers of bio-active compounds stored in the database can be processed automatically. A database which includes these information and suitable for the program ADAM&EVE is particularly preferable embodiment of the present invention.

[0037] By using ordinary three-dimensional databases which include information about element name, three-dimensional coordinates and bonding relation for each constituting atom of the bio-active compounds, it is possible to prepare above-mentioned preferred database suitable for the program ADAM&EVE (ADAM-style database). Since the preparation of ADAM-style database is described in detail, for example, in PCT International Publication WO96/13785, those skilled in the art can easily prepare the database following that procedures or with proper modification and alteration as required. For example, it is possible to assign above-mentioned information automatically after reading the ordinary three-dimensional structure database. If said database does not include the information on three-dimensional coordinates of hydrogen atoms, hydrogen atoms need to be added automatically by calculating their expected position for predicting the most stable structure correctly. If the position of a hydrogen atom cannot be predicted due to a bond rotation, it is desirable to place the hydrogen atom at an extended position in trans form.

[0038] As a preferable method of preparation of the database and addition of the above-mentioned information, an example includes a method in which chemical structures are input by using the ISIS program of MDL company, which is used as a standard for managing market compounds and inhouse compounds, a database is prepared in the form of two-dimensional Molfile of MDL company, structures are transformed automatically to three-dimension by a three-dimensional conversion program, and then above-mentioned information is assigned automatically. However, the database of the present invention is not limited to such prepared with this method.

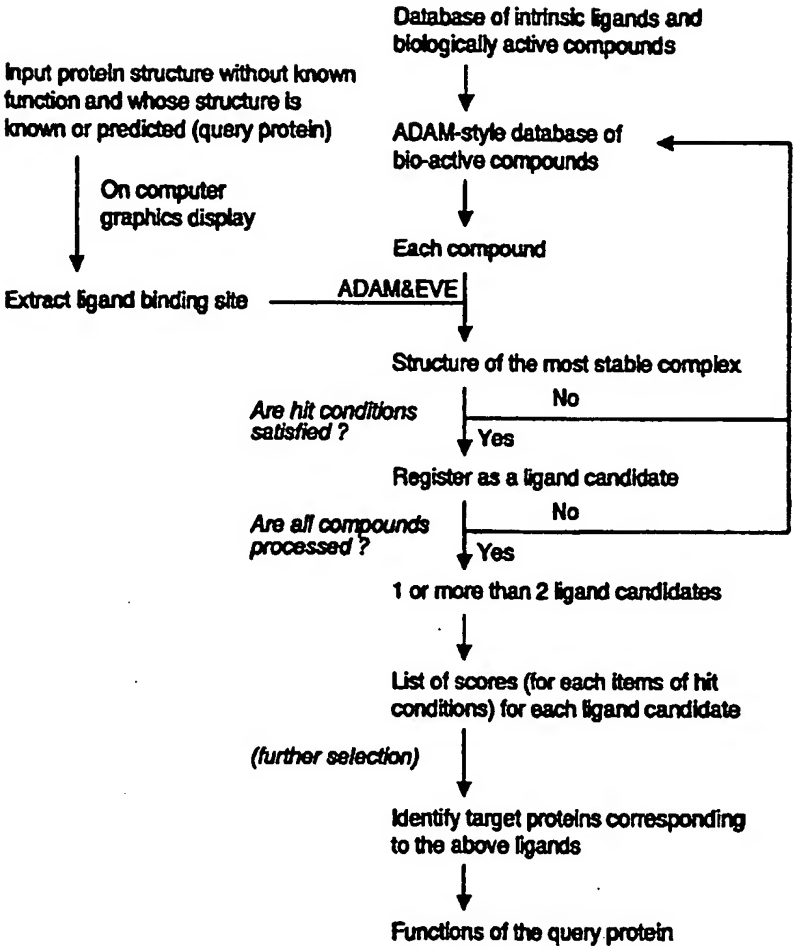
[0039] By selecting hit conditions used in the extraction process of ligand candidates appropriately, it is possible to control the number of the ligand candidates to be extracted. In order to perform the extraction of ligand candidates rapidly and accurately, it is preferable to conduct the extraction process with more than two steps of operation. For example, at the first extraction step, all bio-active compounds with possibility to be a ligand candidate are extracted by applying relatively moderate hit conditions, and at the next extraction step, the most probable one or more most stable complexes can be selected by setting more strict hit condition based on the energy of the complex, number of hydrogen bonds, and other information.

[Prediction of Function of Query Protein]

[0040] Bio-active compounds constituting the most stable complexes that satisfied the hit conditions (ligand candidates) are capable of binding stably to the query protein as ligands. That is, the query protein possesses a ligand binding site identical or analogous to the target protein to which the ligand candidates bind, and accordingly, it is highly probable that the query protein and said target protein have identical or analogous biological functions. It can be also predicted that the role of said bio-active compounds to the query protein is identical or analogous to the role to the target protein (for example, a role of enzyme substrate, receptor substrate and the like). If the target protein is identical for several extracted ligand candidates with different chemical structures, the above-mentioned prediction result is highly reliable.

[0041] For example, if retinoic acid is extracted as a ligand candidate from a database containing various bio-active compounds, it can be predicted that the query protein has a function as retinoid receptor and that retinoic add has a role as an agonist or antagonist to the query protein. Even if identity or analogy to specific target protein cannot be predicted, there is a possibility of predicting the functions of query protein. For example, if a bio-active compound like co-enzyme NADPH which can bind to various biopolymers is extracted as a ligand candidate, it can be predicted that the query protein has either function of oxidation-reduction enzyme utilizing NADPH as co-enzyme or function of enzyme or receptor regulated by NADPH. In other case, if an intrinsic bio-active compound with known bio-activities in organisms but without the knowledge of the target protein is extracted as a ligand candidate, it is probable that the query protein is a novel receptor or enzyme to which the bio-active compound act as an intrinsic ligand.

[0042] As an example of preferred embodiment of prediction methods of the present invention, practical operating procedures using the program ADAM&EVE are shown on the following scheme. However, the methods of the present invention are not limited to the following methods.



[0043] Following the above-mentioned scheme, each step is explained.

1. Select intrinsic ligand compounds in organism such as enzyme substrate, enzymatic products, co-enzymes, signal transducing substances, and hormones, and bio-active compounds whose target proteins are known. Make a database of the present invention by inputting compound names, two-dimensional structures, and other information. Furthermore, input information about the target protein for each of the bio-active compounds.
2. Convert two dimensional structures in the above-mentioned database to three-dimensional structures and create ADAM-style database by adding necessary data automatically.
3. Input three-dimensional structure of a query protein.
4. Specify one or more ligand binding sites (candidate sites) interactively on computer graphics display, and calculate information about three-dimensional grid points, hydrogen bonding, and dummy atoms which is necessary for the calculation by the program ADAM&EVE.
5. Set hit conditions.
6. Select one bio-active compound from the database.
7. Predict the structure of the most stable complex between said bio-active compound and the query protein.
8. Judge whether the structure of the most stable complex described above satisfies the hit conditions.
9. If the hit conditions are satisfied, add said bio-active compound to a ligand candidate group (first extraction group) as a hit and keep its coordinate data and others.
10. Go back to step 6, predict structures of the most stable complexes for other bio-active compounds, and repeat steps 6 through 9 until no more bio-active compound remains to be processed.
11. Concerning the bio-active compounds included in the ligand candidate group (first extraction group), output a list containing the number of compounds, energy value at each complex structure, the number of hydrogen bonds

and others.

12.Reduce the number of bio-active compounds included in the ligand candidate group to a moderate number. As methods for this selection, employ either one of the following methods or combination of more than two methods selected from the followings: a method to select specified numbers of compounds based on ranking; a method to select with more strict hit condition; a method to select interactively on computer graphics display; a method to select with hit condition set by different physical or chemical properties or different computational procedures; and others.

13.Select finally a small numbers of ligand candidates. It is desirable to inspect structure of the complex for each ligand candidates on computer graphics display.

14.Output the classification and biological function of the target protein for each ligand candidate from the database.

15.Predict one or more biological functions for the query protein.

Example

[0044] An example is provided below to describe the present invention more specifically. However, the scope of the present invention is not limited to the example below.

Example 1

[0045] We constructed a small database including bio-active compounds shown in Table 1, and explored capability of binding to a protein with known three-dimensional structure for each bio-active compound contained in the database. Although methods of the present invention can be applied in principle to query proteins without known functions, we used bovine trypsin as a query protein assuming that its function is unknown, and investigated whether or not "nafamostat", which is a trypsin inhibitor, is selected as a ligand candidate. The three-dimensional structure of bovine trypsin and its ligand binding site are shown in figure 1.

Table 1

Bio-active compound	Target biopolymers
Methotrexate	Dihydrofolate reductase
Retinoic acid	Retinoid receptor
Nafamostat	Trypsin
Indomethacin	Cyclooxygenase
Donepezil (E2020)	Acetylcholinesterase
Phorbol ester	Protein kinase C
Morphine	Opioid receptor
Estradiol	Estrogen receptor

[0046] As a result of exploration of the database, nafamostat was selected as a ligand candidate and the compound was shown to bind to the query protein stably. The binding mode of nafamostat to the ligand-binding site is shown in Figure 2. Indomethacin was predicted to have possibility to form a complex, albeit a rather unstable one, and all other compounds was judged not having capability of complex formation (Table 2). From the result of this exploration, the function of the query protein was predicted to be identical or similar to trypsin, which is the target protein of nafamostat.

Table 2

Bio-active compound	Intermolecular Interaction (Kcal/mol)	Number of intermolecular hydrogen bonds
Methotrexate	NA	NA
Retinoic acid	NA	NA

Table 2 (continued)

Bio-active compound	Intermolecular Interaction (Kcal/mol)	Number of intermolecular hydrogen bonds
Nafamostat	-39.6	5
Indomethacin	-29.9	2
Donepezil (E2020)	NA	NA
Phorbol ester	NA	NA
Morphine	NA	NA
Estradiol	NA	NA

Industrial applicability

[0047] According to the methods of the present invention, functions of protein without known functions can be predicted rapidly and accurately. The database of the present invention is useful for conducting the above methods efficiently.

Claims

1. A method of predicting biological functions of query proteins whose steric structures are known or predictable, using a three-dimensional structure database storing one or more bio-active compounds which bind to target proteins with known biological functions, which comprises the steps of:

   (1) extracting bio-active compounds capable of binding to said query protein as ligand candidates from said database based on the capability of complex formation between the query protein and the bio-active compounds; and
   (2) predicting that biological functions of the query protein are identical or similar to the biological functions of the target protein to which said ligand candidates bind.

2. The method according to claim 1, which further comprises the steps of:

   (3) extracting one or more ligand binding sites for the query protein;
   (4) exploring the most stable complex formed with the ligand binding site of the query protein for each bio-active compound included in the database;
   (5) extracting bio-active compounds which satisfy hit conditions preset, based on the stabilities and structural features of the most stable complex;
   (6) extracting further, as required, bio-active compounds from the bio-active compounds extracted in step (5) which satisfy hit conditions different from those in the above-mentioned step (5); and
   (7) predicting that biological functions of the query protein are the identical or similar to the biological functions of the target protein to which said ligand candidates bind, while treating the bio-active compounds capable of complex formation extracted in steps (5) or (6) as ligand candidates.

3. The method according to claim 2, wherein the above-mentioned steps (4) through (6) are performed automatically using the program ADAM&EVE.

4. A method of predicting biological functions of proteins using a three-dimensional structure database storing one or more bio-active compounds which bind to target proteins with known biological functions.

5. A method of predicting biological functions of proteins by exploring shapes and properties of ligand binding site of the query protein using one or more bio-active compounds which bind to target proteins with known biological functions which are stored in a three-dimensional structure database.

6. A method of predicting functions of query proteins by extracting ligand candidates of the query protein from a three-dimensional structure database storing one or more bio-active compounds which bind to target proteins with known biological functions.

7. A three-dimensional structure database storing one or more bio-active compounds which bind to a target protein with known biological functions, which is used in any one of methods according to claims 1 through 6.

8. The database according to claim 7 including information about the target protein for each bio-active compound.

9. A method of predicting biological functions of query proteins whose steric structures are known or predictable, using a three-dimensional structure database storing one or more intrinsic bio-active compounds with known bio-activities in organisms but without the knowledge of the target protein, which comprises the steps of:

(1) extracting bio-active compounds capable of binding to the query protein as ligand candidates from the database based on the capability of complex formation between the query protein and the bio-active compound; and
(8) predicting that biological functions of the query protein concern the bio-activity of said ligand.

10. The method according to claim 9, which further comprises the steps of:

(3) extracting one or more ligand binding sites for the query protein;
(4) exploring the most stable complex formed with the ligand binding site of the query protein for each bio-active compound included in the database;
(5) extracting bio-active compounds which satisfy hit conditions preset, based on the stability and structural features of the moat stable complex;
(6) extracting further, as required, bio-active compounds from the bio-active compounds extracted in step (5) which satisfy hit conditions different from those in the above-mentioned step (5); and
(9) predicting that biological functions of the query protein concern the bio-activities of the ligands, while treating the bio-active compounds capable of complex formation extracted in above-mentioned steps (5) or (6) as ligand candidates.

11. The method according to claim 10, wherein the abovementioned steps (4) through (6) are performed automatically using the program ADAM&EVE.

12. A method of predicting biological functions of query proteins whose steric structures are known or predictable, using a three-dimensional structure database storing one or more intrinsic bio-active compounds with known bio-activities in organisms but without the knowledge of the target protein.

13. A method of predicting biological functions of query proteins by exploring shapes and properties of ligand binding sites of the query protein by using one or more intrinsic bio-active compounds with known bio-activities in organisms but without the knowledge of the target protein which are stored in a three-dimensional structure database.

14. A method of predicting biological functions of query proteins by extracting ligand candidates for the query protein from a three-dimensional structure database storing one or more intrinsic bio-active compounds with known bio-activities in organisms but without the knowledge of the target protein.

15. A three-dimensional structure database storing one or more intrinsic bio-active compounds with known bio-activities in organisms but without the knowledge of the target protein, which is used in any one of the methods according to claims 9 through 13.

Fig.1

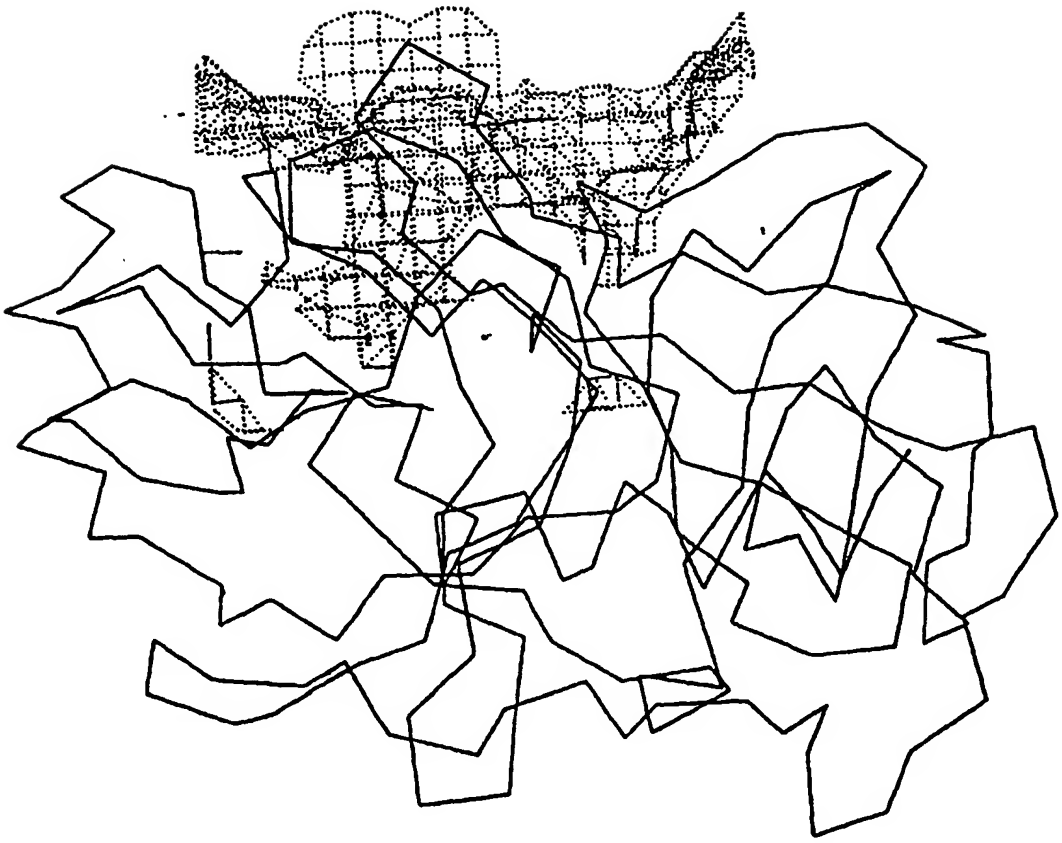
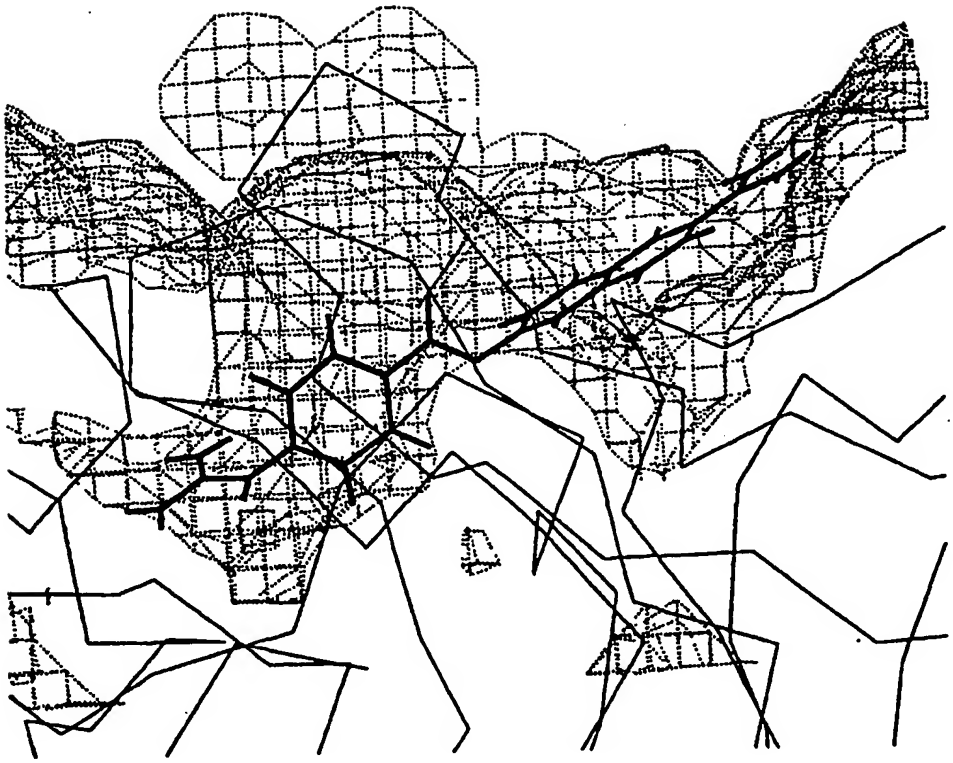


Fig.2



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP98/02986

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[6] C07B61/00, G06F15/40, G06F17/50, G06F17/30, C07K1/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[6] C07B61/00, G06F15/40, G06F17/50, G06F17/30, C07K1/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
BIOSIS PREVIEWS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P, A	WO, 97/24301, A1 (ITAI, Akiko), 10 July, 1997 (10. 07. 97) & AU, 9711528, B	1-6, 9-14
A	WO, 96/13785, A1 (ITAI, Akiko), 9 May, 1996 (09. 05. 96) & EP, 790567, A1	1-6, 9-14
A	WO, 93/20525, A1 (ITAI, Akiko), 14 October, 1993 (14. 10. 93) & EP, 633534, A1 & US, 5642292, A	1-6, 9-14
A	ITAI, A. et al., "Rational Automatic Search Method for Stable Docking Models of Protein and Ligand", J. Mol. Biol., (1994) 243(2) p.310-326	1-6, 9-14
A	TOMIOKA, N. et al., "GREEN: A Program Package for Docking Studies in Rational Drug Design", J. Computer-Aided Mol. Design, (1994) 8(4) p.347-366	1-6, 9-14

[x] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 September, 1998 (29. 09. 98)	13 October, 1998 (13. 10. 98)
Name and mailing address of the ISA/ Japanese Patent Office	Authorized officer
Facsimile No.	Telephone No.

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP98/02986

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	YAMADA, M. et al., "Development of an Efficient Automated Docking Method", Chem. Pharm. Bull., (1993) 41(6) p.1200-1202	1-6, 9-14
A	YAMADA, M. et al., "Application and Evaluation of the Automated Docking Method", Chem. Pharm. Bull., (1993) 41(6) p.1203-1205	1-6, 9-14
A	JP, 2-200641, A (Mochida Pharmaceutical Co., Ltd.), 8 August, 1990 (08. 08. 90) (Family: none)	1-6, 9-14
A	KATCHALSKI-KATZIR, E. et al., "Molecular Surface Recognition Determination of Geometric Fit between Proteins and their Ligands by Correlation Techniques", Proc. Natl. Acad. Sci. USA, (1992) 89(6) p.2195-2199	1-6, 9-14
A	SCHERZ, M.W. "Synthesis and Structure-Activity Relationships of N,N'-di-o-Tolylguanidine Analogues High-Affinity Ligands for the Haloperidol-Sensitive Sigma Receptor", J. Med. Chem., (1990) 33(9) p.2421-2429	1-6, 9-14
A	NISHIBATA, Y. "Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artifitial Lead Generation", Tetrahedron, (1991) 47(43) p.8985-8990	1-6, 9-14
A	GALLOP, M.A. et al., "Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries", J. Med. Chem., (1994) 37(9) p.1233-1251	1-6, 9-14

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP98/02986

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. [x] Claims Nos.: 7, 8, 15

because they relate to subject matter not required to be searched by this Authority, namely:

Claims 7, 8 and 15 pertain to data bases and are therefore considered mere presentations of information. Thus, they relate to a subject matter which this International Searching Authority is not required, under the provisions of Article 17(2)(a)(i) of the PCT and Rule 39.1(iv) of the

2. [ ] Claims Nos.:

because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. [ ] Claims Nos.:

because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. [ ] As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. [ ] As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. [ ] As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. [ ] No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

[ ] The additional search fees were accompanied by the applicant's protest.

[ ] No protest accompanied the payment of additional search fees.

Form PCT/ISA/210 (continuation of first sheet (1)) (July 1992)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP98/02986

Continuation of Box No. I of continuation of first sheet (1)

Regulations under the PCT, to search.

Form PCT/ISA/210 (extra sheet) (July 1992)